DOCUMENT RESUME

ED 221 599 TM 820 637

AUTHOR Masters, Geoff N.
TITLE Looking Ahead in Educational Testing and Assessment.
INSTITUTION Western Australia Education Dept., Perth.
PUB DATE 29 Apr 82
NOTE 54p.; Paper presented to the Western Australian Institute for Educational Research (Perth, Western Australia, April 29, 1982).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Computer Assisted Testing; *Educational Innovation; *Educational Testing; Elementary Secondary Education; *Individual Differences; Instructional Development; *Measurement Techniques; Student Evaluation; Test Construction; Testing Problems; Test Reliability; *Test Use
IDENTIFIERS Tailored Testing

ABSTRACT

Limitations seen in traditional educational procedures are discussed, and three new directions are suggested as being important emphases for testing. A new perception of measurement in which children have positions along particular concept dimensions or "lines" is discussed in terms of traditional testing notions including variables, item banks and test reliability. This new perception could change how educators think about differences among individuals. The direction toward increasing recognition of the individual as the center of testing is suggested. The influence of this shift on test construction; the assessment of performance consistency; and the development of more informative performance reporting and test fairness on an individual basis are discussed. The trend toward making the learning process the focus of testing is the third development considered, and the role of the microcomputer in the integration of learning and testing is examined. A projected computer session on a patient management problem illustrates the implications of the concepts discussed. (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED221599

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Masters

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

RESEARCH BRANCH

EDUCATION DEPARTMENT, WESTERN AUSTRALIA

LOOKING AHEAD IN EDUCATIONAL TESTING AND ASSESSMENT

Geoff N. Masters

Invited address to the Western Australian Institute for Educational Research, Perth, April 29, 1982.

Tonight, at this first meeting of the Western Australian Institute for Educational Research in the new Education Department building, I have titled my talk "Looking Ahead in Educational Testing and Assessment". What I propose to do is to focus not so much on recent developments in psychometrics as on several important new emphases in educational testing which I believe have great promise for improving the quality of testing in our schools. Underlying what I am going to say is my belief that much of what we are doing in educational testing today is misplaced, poorly thought through and of little educational significance. The majority of our efforts in testing and assessment are falling well short of what tests are capable of doing to promote school learning. And worse, tests as they are used in schools today are often a party to lowering the quality of the educational experiences available to children and hindering the development of individual learners.

I will outline what I see as three promising new directions emerging in educational testing. These new directions, when they begin to have an impact on practice, are likely to transform testing as we now know it. Many of our traditional testing notions, procedures and uses will have no place under this new way of thinking, and will need to be rethought or discarded entirely.

I will describe some of the ways in which these new emphases are beginning to appear in practice and outline some of the possibilities which they raise. Some of these possibilities are being explored, some exist only as ideas, and still others we

are not yet able to handle satisfactorily psychometrically. All, however, have important implications for testing and learning in our schools.

Many of the shortcomings of testing and asssessment today are, I believe, attributable to the fact that so much of what we do in the testing area is grounded in tradition -- and, in particular, in the psychological tradition. The thinking of men like Galton and Terman continues to have a profound influence not only on our testing procedures, but also on the uses to which tests are put in our schools, and even on the ways in which we think about school children.

In 1869 Francis Galton envisaged tests being used as "shafts" to assess the quality of the school's "raw material", much as the quality of a mineral deposit might be assessed. Half a centuty later, Galton's vision had become reality. By the 1920s standardized testing had been swept into the schools. Suddenly millions of children were being administered standardized tests to assess their "mental capacities". Test results were being used to classify children and to develop curricula appropriate to their "natural endowments". In 1922, Lewis Terman, as chairman of a national subcommittee on the role of testing in re-organizing American elementary schools, wrote:

> "There are few if any more significant events in modern educational history than the developments which have recently taken place in methods of <u>mental measurement</u>. The importance of this new psychological tool for the improvement of school administration has been recognized everywhere with a promptness which is hardly less than amazing ... Intelligence tests have demonstrated the great extent and frequency of individual differences in the mental ability of unselected

> school children, and common sense tells us how necessary it is to take such differences into account in the framing of curricula and methods, in the classification of children for instruction, and in their educational and vocational guidance. Standardized tests of the school's raw material can no more be dispensed with than standardized tests in agriculture, manufacturing or medicine". (Terman, 1922, 1-3)

Today this use of tests to select, classify and stream on the basis of "capacity" or "aptitude", for the purposes of school administration, continues to account for a large share of our efforts in school testing. However, much of this psychological legacy is, I believe, out of step with current educational thinking and inappropriate in today's schools.

The psychological tradition has influenced not only the uses to which tests are put in schools, but also our testing procedures. The testing enterprise as it exists today, is encased in a language and in a set of procedures which we have almost come to regard as sacred, and have even begun to believe. "The normal curve", "percentiles", "Z-scores", "test reliability", and "test validity" are all a part of the language and the procedures which we not only accept, but which continue to shape and to direct our testing efforts in schools. If in doubt, we have a handbook of "standards" to which we can turn -- the joint policy statement of The American Psychological Association, The American Educational Research Association and The National Council on Measurement in Education -- in which appropriate procedures for test constructors and users are carefully enumerated. Featured in pride of place on the cover of this handbook is the most sacred of all psychological testing legacies -- the normal curve.

Before looking ahead I will mention briefly what I see as several of the ways in which tests can, and in some cases are, lowering the quality of education available to individual learners in our schools. Of course, there has been no shortage of criticism of the use of tests in schools. I do not wish to merely add to this. Rather, I mention these few unfortunate consequences and uses of tests in our schools because they continue to receive a large share of the testing profession's attention, and because they bear directly on what I have to say later.

Narrowing and Prescribing

Perhaps the most common criticism of standardized tests is that they have a narrowing influence on what is taught in classrooms. Externally developed standardized tests for selection or for the comparison of standards typically contain 30 or 40 carefully constructed and trialled items. In a sense, these few items completely define the construct measured by the test and so, must be carefully protected. If the items on a verbal ability test became public, all that is necessary for a child to make a high verbal ability score is for teachers to teach the few tasks which appear on the test. A certain psychologist once commented that IQ is what an IQ test measures. For this he was much criticized, but in fact, IQ is exactly what an IQ test measures. If you wish to know what IQ is, the only way you are going to find out is to examine the few items on an IQ test. And, as numerous popular books point out, all that is necessary to "Improve Your IQ" or to "Improve your SAT Score" is to improve your test taking skills and to master the few types of

tasks which appear on these tests. When teachers teach to the limited content of standardized tests, particularly tests which assess only recall or recognition, they lower the range and quality of the learning experiences available to children.

The possibility that standardized tests will have a narrowing and prescriptive influence on the curriculum is, of course, well recognized, and is generally accepted as the lesser of two evils. When selection for special programs or the comparison of standards is required, standardized tests seem preferable to haphazard testing at the school level.

Lowering Expectations and Denying Opportunities

A second way in which tests are having an undesirable influence on what happens in schools is through the lowering of teachers expectations of particular children. In many of our testing and assessment efforts, tests are a party to entrenching the notion that children have marked differences in aptitude or capacity to succeed in school. We use tests to sort children into groups -- those from whom much is expected, and those from whom little is expected. Tests not only assist in this sorting, but also give it an air of legitimacy.

This use of tests has a long tradition. It dates back to Galton's vision of sinking "shafts" to assess the quality of the school's "raw material". Galton believed that tests of capacity could be used to sort children into their "correct" classes and future occupations. He considered it important that students recognize as early as possible their own natural intellectual "endowment". "When a person becomes aware of his own intellectual capacity", he wrote, "he is no longer tormented into

hopeless efforts by the fallacious promptings of overweening vanity, but he limits his undertakings to matters below the level of his reach, and finds true moral repose in an honest conviction that he is engaged in as much good work as his nature has rendered him capable of performing" (Galton, 1892, 15).

Galton proceeded to show how areas under the normal curve could be used to define groups of children which could then be assigned labels -- from the "idiots and imbeciles", the "feebleminded" and the "mediocre" through to the "gifted and geniuses". Noting that his grouping placed most individuals in one of the central groups, and only a few in the extreme groups, he commented:

> "the vast abundance of mediocrity, is no accident, but follows of necessity from the very nature of these things."

Galton thus showed how the normal curve could be used to justify expecting mediocre performances of the vast majority of children. This remains the most important role of the normal curve in education today.

As we read Galton today we find much of what he had to say either offensive or amusing. We no longer use words like "idiots" and "imbeciles" as scientific terms. We no longer believe that the races of mankind can be ordered on the basis of intellectual capacity "from caucasian to the lowest savage (the Australian aborigine)". We find his attempt to equate the "second highest class of dog" with the "second lowest class of human" offensive. We instantly recognize his language as sexist, and we find his unswerving faith in the relation between head size and intellectual capacity amusing.

And yet, Galton's views on intellectual capacity or aptitude underlie so much of what we do in educational testing today. Our "intelligence" and "aptitude" tests are legacies of the psychological tradition in testing, descended directly from Galton's attempts to find students' "levels of reach". We use tests to tell us what to expect of children in our schools, and worse still, to tell children what to expect of themselves. And we use tests to assign children to streams where they can limit their undertakings to matters below the level of their reach. As Benjamin Bloom has noted,

> "Most researchers and educators have interpreted the relations between intelligence and aptitute test scores and achievement as indicating that these characteristics (intelligence and aptitude) determine the individual's potential for learning. They use them as a basis for making long-term decisions about selection, streaming, and even about the types of school programs to be assigned individuals. All too frequently, intelligence and aptitude scores determine the individual's opportunities for further education, support, and encouragement in the school and even the qualities of the interaction between the teacher and the student." (Bloom, 1981, 4-5).

When tests are used to assign children to groups (usually based on areas under the normal curve), they invariably lock children in to the groups to which they have been assigned. Even when tests of achievement are used to assign children, the resulting groupings are invariably interpreted as reflections of capacity to learn:

> "In the achievement certificate system as we have perceived it, there is a strong tendency for students allotted to lower-course levels to remain in those levels, confirmed there by the syllabuses, the teaching, and the lack of pressure both in school and in homework." (Priest, 1981, 40)

My own belief is that we will one day come to view Galton's levels of intellectual "capacity" with about as much credence as we now view his claims about racial differences in natural capacity, or his belief in the relation between headsize and capacity to learn. This will come about, I believe through an improved understanding of the factors underlying high achievement. Bloom and Sosniak's (1981) study of exceptionally high achievers, for example, suggests that among the factors underlying the exceptional achievement of the individuals they studied were: very early commencement and dedication to the field, exceptional home support and family commitment, individualized tutoring, and perhaps especially important, exceptionally high expectations of and confidence in the individual concerned:

> "when the child is learning poorly it is assumed that the teacher is at fault and the parents seek another teacher." (Bloom and Sosniak, 1981, 90).

While we continue to use tests to set our expectations of individual learners, and to assign them to streams in which these expectations can be realized, tests continue to be a party to hindering the development of children in our schools.

Setting Minimum Standards

Finally, tests are being used to define minimum levels of school achievement. This particular use of testing has become widespread in the last few years as a result of concern over standards, particularly standards of achievement in literacy and numeracy. Following the introduction of legislation requiring evidence of "minimal competency" for high school graduation, test constructors have been called upon to develop

tests capable of assessing whether or not students possess the minimum literacy and numeracy skills necessary to function in society. In this state too, there have been calls for testing efforts to specify and to assess minimum achievement:

> "Comments from some critics of our schools have been to the effect that there should be a specification of the essential minimum skills that schools should develop in their students, and that tests should be devised for their measurement" (Priest, 1981, 41).

This use of tests to define minimum standards of performance for graduation invites minimum achievement. Once the arbitrary boundary between competence and incompetence has been set, this becomes something to be aimed for, and the incentive for high achievement is removed. I think that there is some evidence to suggest that school systems which strive for minimal competency among their students are doomed to succeed.

Frequently our efforts in testing are seen as legitimizing the purposes for which tests are developed. Scores on "intelligence" and "aptitude" tests are interpreted as reflections of childrens' capacities to succeed in school. The observation that children make different scores on these tests is taken as evidence that children really do differ in capacity. The use of tests to "identify" children of particular "capacities" (e.g., "gifted" children) and to assign them to special streams appears to be vindicated when, as a result of this streaming, our expectations of these children are fulfilled. And the observation that some children score below and others score above some arbitrary criterion on a minimal competency test gives an air of legitimacy to the notion that there exist "competent" and "incompetent" children.

Much of what we are doing in educational testing and assessment today, both in terms of our testing procedures and in terms of the uses to which tests are put, has been inherited from the psychological testing tradition. Many of the trappings of this tradition are, I believe, inappropriate in educational measurement and are lowering the quality and utility of our testing efforts in schools.

Increasingly, I believe, those of us who see ourselves as specialists in testing and assessment will be faced with questions about the uses to which our tests are put. We have a responsibility to the children who take our tests to ensure not only that testing methods and procedures are as psychometrically sound and defensible as possible, but also that the uses to which our tests are put are as educationally defensible as we can make them.

A NEW CONCEPTION OF MEASUREMENT

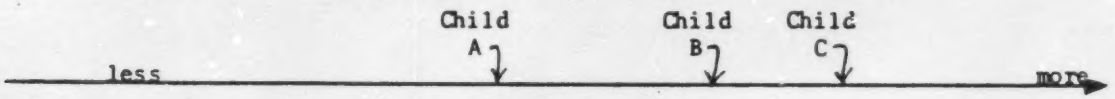
Perhaps the most significant development which is taking place in educational testing today is the emergence of a new conception of what measurement is about. This development is bringing with it a change in the way we think about testing. Under this new way of thinking, many traditional testing notions and procedures like the normal curve, percentiles, z-scores, stanines, test norms, test reliability, test validity, item discrimination and random sampling play either a different role, or no role at all.

Traditionally, formal educational testing has been synonymous with the standardized test. This test contains a fixed set of carefully constructed and trialled items which, because they completely define the construct measured by the test, must be carefully protected. The test is developed externally, usually by some testing agency. Evidence is accumulated for the test's reliability and validity, and a table of norms is compiled so that test scores can be interpreted as percentiles with respect to some group. Underlying the use of the standardized test is the notion that valid measurement is ensured, to the extent that it can be, by careful test construction.

The Idea of a Variable

In contrast, fundamental to the new conception of measurement is the notion that children have positions along particular dimensions or "lines". These lines are not real. They are ideas. But they are powerful ideas which pervade our thinking about differences among individuals. We invent dimensions so that we can focus our attention on one aspect of a person's development at a time. A single dimension or "variable" is implied whenever we say that one person is achieving at a higher level than another or that one person is more knowledgable than another. The notion of a single dimension is also implicit in every attempt to report a single score for a person on a test.

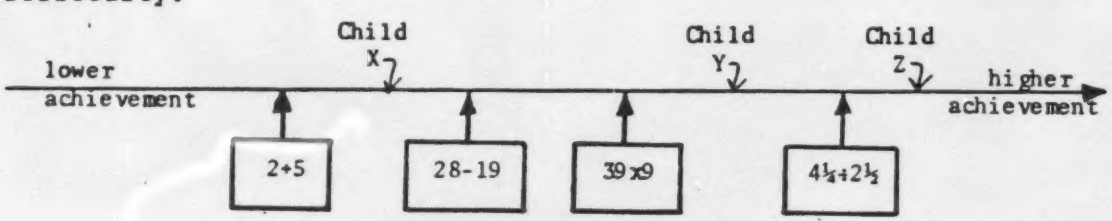
The very idea of "reading proficiency", for example, implies one line along which children can be positioned. At any given time the poorer readers are on the left of this line and the better readers are on the right. A child's progress in reading is monitored by tracking his movement along this line.



"reading proficiency"

Whether it is useful to imagine that children have positions on a line of increasing "reading proficiency" depends on the extent to which this idea is supported in practice. The idea of a reading variable will be supported only if independent attempts to assess reading point to more or less the same ordering of children. If different reading tasks produce quite different orderings of children, then our attempts to define a line will be unsuccessful and we will be unable to draw any meaningful inferences about "poorer" and "better" readers.

When tasks or items work together to support the concept of a single dimension, the tasks themselves can be ordered from easy to hard on the basis of how often they are successfully performed. This ordering of tasks provides the operational definition of the variable. A mathematics computation variable, for example, is supported by mathematics tasks of increasing difficulty.



"Mathematics Achievement"

These tasks give the variable meaning. They enable us to see what it means for Child Z to be achieving at a "higher" level than Child X in terms of the types of mathematics operations each child has mastered. The notion of a single variable, marked out by items which work together to support this idea, is central to this new conception of measurement. The process of estimating the positions of items along a line is called "calibration". The process of estimating positions for persons is called "measurement".

The idea of a variable provides an opportunity to measure one aspect of a person's development. But this will rarely be everything we want to know about a person. In addition to measuring a person's reading comprehension we may also wish to assess how much he <u>enjoys</u> reading and, if we are interested in assessing literacy in general, how well he <u>writes</u>. A single variable can be used to measure only one dimension of a person's achievement but, as Thurstone noted, this is a fundamental feature of all measuring systems.

> "The measurement of any object or entity describes only one attribute of the object measured. This is a universal characteristic of all measurement. When the height of a table is measured, the whole table has not been described, but only that attribute which has been measured" (1931, 257).

Of course it could be argued that the idea of a single "reading" variable is an oversimplification; that reading is a collection of skills, each of which could be assessed. This is true. All variables are simplifications. Whether we choose to think of reading as a single variable or as several variables depends on our purpose. If we see no reason why the most skilled

persons in one area of reading should not also be the most skilled persons in other areas, and if the items we construct to assess these skills can be shown empirically to work together to support the idea of a single reading variable, then we will probably report one reading score for each child. If, on the other hand, we identify particular reading skills and teach these skills separately, then we may want separate measures on each skill. Notice, however, that if we report scores on two or more skills and then add these to obtain a "total" reading score for each child, we fall back on the notion of a single line of increasing reading proficiency.

Items as Replications

Under this new conception of measurement, items are viewed not as fixed and immutable, but as transient and expendable. They are examples of the variable. Each item helps to define that level of the variable at which it is calibrated, but that level could supposedly be defined equally well by one of a large number of other examples. The item 1/2 + 3/4 = ?, for example, could probably be replaced on a test by the item 1/2 + 2/3 = ? or 1/3 + 1/4 = ?, both of which should have very similar positions on a line of increasing mathematics achievement.

An implication of this view of items as "examples" is that instead of administering a secure standardized test to measure a person's achievement, a set of calibrated items can be selected from a bank. Which particular items are selected is unimportant, provided that they are of appropriate content and difficulty for

the persons taking them. And, as there is no limit to the number of examples which could be accumulated, teachers can propose their own items for inclusion in a bank. These items might result from attempts to replicate items of a particular difficulty, to fill "gaps" in a variable where there are few items, or to cover other content areas (e.g., the addition of magnetism and electricity items to a science bank containing primarily chemistry and mechanics items). However, the addition of new items to a bank is contingent upon empirical evidence that they are consistent with the single line of increasing achievement defined by the existing bank items.

"Test-free" Measurement

An item bank is potentially a powerful measuring system. But its potential can be realized only if it frees us to select any appropriate set of items from a bank to measure a person. If the measure of a person's achievement fluctuates wildly, depending upon which particular items we choose to give him, then we will be unable to draw any meaningful inference about his position on the variable the items are supposed to define.

The possibility of choosing any set of calibrated items from a bank and obtaining more or less the same measure for a person is called "test-free" measurement. When we measure a person's height, weight or temperature, we expect the measure we make to have a meaning which transcends the instrument we use to obtain it. We expect to obtain about the same measure of a person's height whether we use a wooden rule or a steel tape. This is an example of instrument-free measurement.

Traditional test scores are not instrument-free. A person's score on a test is a number which depends not only upon his level of achievement, but also upon the number of items in the test and on the difficulties of those items. The scores obtained when persons take tests which differ in length or difficulty cannot be used directly to position persons along a line. They must first be adjusted for the number and difficulties of the items used to obtain them. Here again, traditional indices of item difficulty will not suffice. The traditional index of item difficulty, the proportion of persons in some group who get the item wrong, depends on the abilities of the persons in that particular group. Because it is rarely possible to give every item in a bank to the same group of persons, items are usually calibrated on different samples. Sample-dependent indices of item difficulty, like proportion correct, suffer from the same problem as test scores. Before they can be used to mark out a line of increasing achievement, they must first be adjusted for the abilities of the persons in their calibrating samples.

Of course, the possibility of making "test-free" measures does not mean that we would select just any items to make a test. Instead, we will select items which we believe are of an appropriate difficulty for the children we wish to measure. We will also select items appropriate to the content areas that have been covered. A science bank, for example, may contain items in chemistry, genetics, mechanics and electricity, all calibrated along a single line of increasing science achievement. From this bank teachers can select items in the topics they have taught.

Because all items are calibrated along the same science variable, comparisons can then be made between students who cover different science topics.

Test Reliability and Validity

This new conception of measurement also has implications for the traditional notions of test "reliability" and "validity". In latent trait theory, the first question to be asked of a set of test data is whether or not it supports the idea of a single line along which persons and items have unique positions. Whether it is valid to report a single position estimate for an item depends on the extent to which responses to that item are consistent with the idea of it having a single difficulty. If many low-scoring persons pass an item, and many high-scoring persons fail, then the validity of that item's estimated position on the line is suspect. In this way, the internal validity of a test can be assessed on an item-by-item basis. An item is valid to the extent that it works with the other items to define a single line of increasing achievement.

The "construct validity" of a test is assessed by examining the empirical ordering of the items along a line and asking whether this ordering makes sense. Often it will be useful to ask subject area specialists to attempt to order items by difficulty before a test is administered. To the extent that this expected order is reflected in the obtained difficulty order, the construct validity of the test is supported.

In addition to assessing the consistency of responses to each item, it is also possible to assess the validity of reporting a single position estimate for each person. Only if a person's responses are consistent with the difficulty ordering of the items in a test will the reporting of a single measure for that person have any validity. If a person fails many easy items and succeeds on many difficult items, then the validity of their estimated position on the variable is suspect.

In traditional test theory "reliability" is conceived of as an index of the "consistency" of test scores. But the magnitude of the reliability coefficient depends not only on the precision with which the test measures (the average test error variance), but also on the variance of the sample true scores. When reliability is estimated as a correlation (e.g., between split-halves), its magnitude also depends on the extent to which the items in the test work together to define a single variable.

These three components of traditional reliability, error of measurement, sample variance and internal consistency, are estimated separately in latent trait theory. Latent trait models provide an estimate of the error of measurement, not for the test as a whole, but for each ability estimate. This is important because the error of measurement can be quite different at different levels of ability. Unlike the error of measurement derived from reliability coefficients, this error is not influenced by sample variance or by internal consistency.

MEASURING INDIVIDUALS

A second emerging emphasis which I believe has far-reaching implications for the quality and relevance of educational testing is the increasing recognition of the individual as the center of our testing efforts. At first glance the significance of this shift in emphasis may not be apparent. After all, tests have always been taken by individuals, and test scores have always been reported on an individual basis. However, I think that the history of standardized testing shows not only that the measurement of individuals has been of secondary importance, but also that the procedures of the traditional testing movement have often been designed to ignore and to deny the individuality of testees.

The history of standardized testing has been a history not of person measurement, but of test construction. The emphasis of this tradition has been on the construction of tests. To this end, the concern has been with estimating test and item characteristics. These include item statistics like difficulty, discrimination, item-total correlation, point biserials, and tetrachoric correlations, and test characteristics like test-retest, parallel forms and split-half reliabilities, KR-20 and the various indices of internal and external validity. The literature is full of imaginative ways of assessing the quality of items and of tests.

Underlying this tradition is the notion that once a test has been carefully constructed, all that can be done has been done to ensure valid measurement. All that remains is to obtain norms

for the test in terms of which individual scores can then be interpreted.

In the test construction tradition it is considered essential to check on the internal consistency of a test and, in looking at an individual item, to ask whether persons with high test scores have been more successful on that item than persons with low test scores. If this general expectation is not met, then the item is considered flawed, and rejected from the test. Interestingly, although the various statistics for making such checks on individual items have been available for more than half a century, this same question has almost never been asked of the consistency of an individual's responses to the items on a test: Did this person get easy items right and hard items wrong? And if not, can we conclude anything meaningful about his/her score? Instead, the procedure has been to simply report a score of 30/40, however it is made, and to refer it to a table of norms.

More recently, this tradition has been applied to the construction of item banks. The notion is the same -- develop a pool of items, trial them, prune out items which are not functioning as well as expected then, with the task completed, turn the "bank" over to the schools. Once again it is believed, mistakenly, that a carefully constructed pool of items is sufficient to ensure valid and comparable measures of achievement.

This test construction tradition also continues in much of item response theory, or item characteristic curve theory, as it is sometimes called. Here, the very names given to this

activity betray its emphasis. The emphasis is on estimating item parameters, with measurement a secondary concern. And, as in all branches of the test construction tradition, item and test characteristics proliferate. Item difficulty, item discrimination, lower asymptote, upper asymptote and item fit have all been proposed for study in item response theory.

The test construction tradition has not been much interested in the performances of individuals. It has avoided developing statistics and procedures to study individuals, and has preferred instead to invent and to study item and test characteristics. In fact, when possible, the approach has been to ignore individual testees and to think of them instead as groups characterizable by means and standard deviations. Indeed, some item response theory methods require for their very operation that persons be seen not as individuals but as a group for which some distribution of abilities can be assumed. And, in the tradition of test construction with its emphasis on estimating item parameters, calibration frequently becomes an experiment for which "subjects" are sampled at random from some population.

Taking this one step further, if we are prepared to think of persons not as individuals, but as random events then we can introduce into our testing procedures the distribution of random events, or normal curve. With this further step, test scores can be transformed to z-scores, T-scores, stanines or IQ-equivalents, and from here, as Galton showed, it is a small step to dividing children into groups to which labels can then be attached.

It may not be going too far to suggest that the normal curve has been the scurge of educational testing in this century. It has been used not only to deny the individuality of children, but also to set our expectations of children and to assign them labels which, even when based on measures of achievement, are interpreted as reflections of capacity to learn. We have come to believe, with Galton, that the vast abundance of mediocrity is no accident, but is to be expected, and follows directly from the shape of the normal curve. And so, the normal curve remains central to so much of what we do in educational testing today. Almost every book on test theory begins with some nonsense about scales of measurement and then launches into a detailed discussion of the normal curve, often showing how some attribute like height is approximately normally distributed in a random sample of individuals, but never pointing out that the assumption of normality plays absolutely no role in the measurement of such attributes.

Finally, in much of our testing and assessment work we ignore individual performances quite deliberately, as for example, in attempts to monitor group standards. Under political pressure, this use of tests has come to demand a large share of the testing profession's time and energy in recent years. The large and expensive national monitoring projects in the U.S. (the National Assessment of Educational Performance) and the U.K. (the Assessment of Performance Unit) are the most prominent examples of this group monitoring. But in this country too, considerable effort is being put into monitoring group standards of achievement.

Because these testing efforts look only at groups they are inevitably of little educational significance, and in fact to date, appear to have produced very few policy implications at all. At worst these studies have produced conclusions like:

> In the past five years the average performance of Year 8 children on a standardized test of mathematics achievement has declined by two score points.

The educational significance of such "declines" is rarely considered and, even supposing that is makes sense to talk about the performance of the "average" child, there is rarely any attempt to interpret this two-point decline in terms of what he/she can now not do that the "average" child could do five years ago.

Only slightly better are attempts at "criterion-referenced" interpretations of standards. These interpretations are based on randomly selected samples of children and test results are reported item by item. These studies lead to conclusions like:

> In the past five years the percentage of 17 year olds who know that salt is more effective than water for putting out a grease fire has dropped from 67% to 62%.

Again, the implications of such conclusions for educational practice are obscure, to say the least.

In pointing this out I am not suggesting that it is never useful to monitor group performance. What I am saying is that this use of tests has very little educational significance, and is probably receiving far more attention from the testing profession and far more funding than it justifies. Certainly, as a parent, I am not much interested in whether the "average" child in my son's age cohort is scoring two points lower or two

points higher than the "average" child of that age did five or ten years ago. Of far more interest to me would be information about what my son can now do that he could not do three months ago. What learning experiences is he now ready for? And, above all, what are the gaps in his understanding that threaten to impede his progress through the next unit of work? It is the development of techniques and procedures for answering questions like these which I believe are most deserving of our future efforts in testing and assessment.

I turn now to consider some of the ways in which an emerging concern for <u>individuals</u> is beginning to be manifested in testing practice. An important impetus in this trend is the work of Georg Rasch, and his concern for objectivity in testing. Rasch's interest in measuring individuals began while he was in London in 1934 studying data on crab shell structure. He became convinced that it was possible to study growth on an individual as well as on a population basis. In 1979 he reflected:

> "It meant quite a lot to me to realize the meaning and importance of dealing with individuals and not with demography. Later I realized that test psychologists were not dealing with the testing of individuals, but ... were studying how traits, such as intelligence, were distributed in populations. They were making demographic studies and not studies of individuals" (Rasch, 1980 xii).

Rasch's measurement models are an important break with the approaches and methods of the test construction tradition which treat a person's ability as a random event in a distribution of abilities. Much of the emerging recognition of individuals in testing and assessment stems from his work.

Tailored Testing

One way in which this trend is being manifested is in the growing concern for the construction and use of tests which are appropriate for the individuals who take them. The standardized testing tradition has not been particularly concerned with the development of tests which are of suitable content and difficult for individual testees. This has been a frequent criticism of standardized tests of achievement:

> "It is held by some teachers that there remains an undue restriction on their freedom to organize their work according to their professional judgement. If the subject matter and skills have not been covered at the time of testing because a teacher has re-arranged the sequence of work, ... achievement assessments may be unfairly influenced... Often the lowest achieving students have not even covered much of the content and skills on which the test items are based, and they know they are doomed to failure." (Priest, 1981, 29)

Concern over providing tests of appropriate content and difficulty is being manifested at three levels. At one level is the attempt to develop a variety of test forms in a school subject, each covering some combination of the alternative topics which a student may have covered. This attempt to provide tests more suited to the content material covered lowers the level of restriction placed on teachers' decisions about the curriculum and provides an opportunity to make a fairer assessment of an individual's achievement than is possible with a single standardized test. Of course, if meaningful comparisons of performances on different tests are to be made, a methodology capable of linking these tests and of supporting such a measurement system must be used.

At a second level, attempts are being made to develop tests which are not only of appropriate content, but also of appropriate difficulty for the persons taking them. Tests which are tailored to the abilities of groups of children not only maximize the precision of measurement, and so provide more information about a child's ability, but also reduce the likelihood of undesirable testing behaviors like carelessness and guessing. They should also provide children with a greater sense of accomplishment and a more positive view of themselves as learners.

At a third level, attempts are being made to provide tests of appropriate difficulty and content for individuals. While this has been difficult in the past because of the problem of assembling different paper and pencil tests for different individuals, the introduction of microcomputers into test administration and scoring makes the individualization of testing relatively straightforward. By the turn of the century, much of our current paper and pencil testing will be replaced by tests administered and scored by microcomputers. Items will appear one at a time on a screen, and the child will enter his answers on a keyboard. The items will be stored in a bank ready for retrieval. The particular items an individual takes may be specified ahead of time by the teacher, they may be drawn at random from a specified content X difficulty grid appropriate for that individual, or they may be selected automatically on the basis of the child's performances on the items that they have already taken. If a child has succeeded on several easy items,

then a more difficult item of the same type might be automatically selected. In this way, the matching of test difficulty to an individual's ability could be made almost ideal.

Analyzing Consistency

A second way in which this trend towards the greater recognition of individuals is being manifested in our testing efforts is in an increasing concern with assessing the consistency of an individual's test performances, and hence the fairness of reporting and interpreting a person's test score as a measure of their achievement. The importance of examining the consistency of an individual's responses was recognized by Thurstone in the 1920s. In developing attitude scales he looked at the response patterns of individuals to the items on questionnaires and assessed their consistency with the ordering of the questionnaire items. "Inconsistent" response records, as he called them, were rejected from his attempts to scale questionnaire items along a continuum of increasing attitude. However, Thurstone's work has found no place in the procedures of standardized testing. In the mainstream of standardized testing there has been virtually no attempt to look at the consistency of individual response patterns.

A first, and rather dramatic, illustration of the necessity of looking at the consistency of an individual's test performances is shown in Table 1. Table 1 shows the scores of one candidate on six booklets of a professional certification examination. These scores have been converted to ability

TABLE 1
PERFORMANCES OF ONE CANDIDATE ON SIX PARALLEL TESTS

TEST BOOKLET	SCORE	ABILITY	ERROR	ABILITY PLOT -2 -1 0 1 2	WITHIN FIT-T
1	29	-1.54	.23	-X-	5.8
2	96	.42	.18	-X-	.7
3	81	.45	.18	-X-	.3
4	82	.34	.19	-X-	1.1
5	90	.59	.19	-X-	1.6
6	74	.49	.20	-X-	.3

From Wright, B.D. and Masters, G.N. Rating Scale Analysis, MESA Press, Chicago, 1982.

estimates with an associated error of measurement. Unlike the scores, which reflect the number and difficulties of the items in these booklets, the ability estimates are expressed in the same metric and so, can be compared directly.

A plot of the six ability estimates is shown on the right of the table, and on the far right is an index describing the consistency of the person's performances on the items in each booklet. The higher this index, the less consistent the person's performances (that is, the smaller the tendency for this person to get easy items right and hard items wrong). The ability plot shows that five of the booklets point to much the same ability estimate for this candidate. The first booklet, however, points to a much lower ability. This person's performances on the items in this first booklet are also internally very inconsistent, as indicated by the large positive fit index (5.8).

The anomalous score on Booklet 1 was readily explained upon further investigation. This candidate's answer sheet to Booklet 1 had been crumpled as it went through the optical scanner, and as a result, the record had been completely misread. The simple addition of this candidate's scores on the six booklets to obtain a total score which could then be compared to some certification criterion would not have exposed this problem, and the consequences for this candidate could have been disastrous. This is an extreme example, both in the way in which it occurred, and in its possible consequences. But, it illustrates the necessity of checking to see that it makes sense to combine performances to obtain a single score for a person. Unless an

individual's responses are consistent, they cannot be used to obtain a meaningful score which can then be compared with other test scores.

A second illustration of the need to look at individual response patterns is shown in Table 2. Table 2 shows the responses of four young children to 25 items on a questionnaire constructed to measure liking for science. Three response alternatives were provided: "Dislike" (Ø), "Not Sure" (1) and "Like" (2). The 25 questionnaire items are shown at the top of the table and are ordered from the activity easiest to like (going on a picnic), on the left, to the activity hardest to like (finding old cans), on the right.

The responses of Child 8 are reasonably consistent with the difficulty ordering of these 25 activities. This child likes the easiest to like activities, is unsure about three, and dislikes the activities hardest to like. As a result, this child's score of 33 can be meaningfully used to position this child on the liking-for-science continuum defined by these 25 items. We can use his score to infer the types of activities he is likely to enjoy. He will certainly enjoy outings. But he also likes finding out about plants and animals from books and encyclopaedias.

The responses of Child 71, on the other hand, are much less orderly. Child 71 obtained the same score as Child 8, and even made his score in the same way -- by responding "Like" to 15 activities, "Not Sure" to three, and "Dislike" to eight.

TABLE 2

Child	going on a picnic	going to the zoo	going to a museum	listening to bird sing	growing a garden	finding where animal lives	watching bird make nest	reading books on animals	stories of animals	watching birds	finding what flowers live on	finding what animals eat	watching what animals eat	looking up in encyclopedia	looking at pictures of plants	reading books on plants	talking about plants	making a map	learning names of weeds	watching animal move	looking in cracks for animals	watching grass change color	watching bugs	watching a rat	finding old cans	Score	frequency of response 0	1	2
8	2	2	2	2	2	2	2	2	2	2	2	2	1	2	0	1	2	0	2	1	0	0	0	0	0	33	7	3	15
71	2	(0)	2	(1)	(0)	2	2	2	2	2	0	2	2	2	0	0	1	2	0	2	0	1	(2)	(2)	(2)	33	7	3	15
62	2	2	2	2	2	2	2	2	2	2	0	0	2	0	2	2	0	0	2	0	0	0	0	0	0	28	11	0	14
52	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	27	0	23	2

EASY TO LIKE

HARD TO LIKE

0 = "DISLIKE"

1 = "NOT SURE"

2 = "LIKE"

However, Child 71's pattern of responses leaves us puzzled about his liking for science. He dislikes going to zoos, listening to birds and growing gardens, but likes watching bugs and rats and finding old cans. His responses do not enable us to draw a meaningful inference about his position on this line of increasing liking for science.

Children 62 and 52 raise another issue in the analysis of response patterns. Both patterns are orderly, with 2's on the left and Ø's on the right. But Child 62 did not use the middle alternative at all. His complete avoidance of the middle category is atypical of children in this group, and raises a question about the validity of comparing his score with the scores made by the majority of these children. Child 52, with a very similar score of 27, made his score in an equally orderly but surprising way. This child responded "Not sure" to all but two of these 25 activities. This leaves us puzzled about the interpretability of his score too.

Had we simply reported scores of 27, 28 and 33 for these four children and tried to compare them, as would usually be done with such a questionnaire, we would have been unable to make any meaningful comparisons because the scores were made in quite different ways. Before scores on a test or questionnaire can be compared, they must arise from response records which are consistent with the ordering of the test or questionnaire items.

Individual Reports

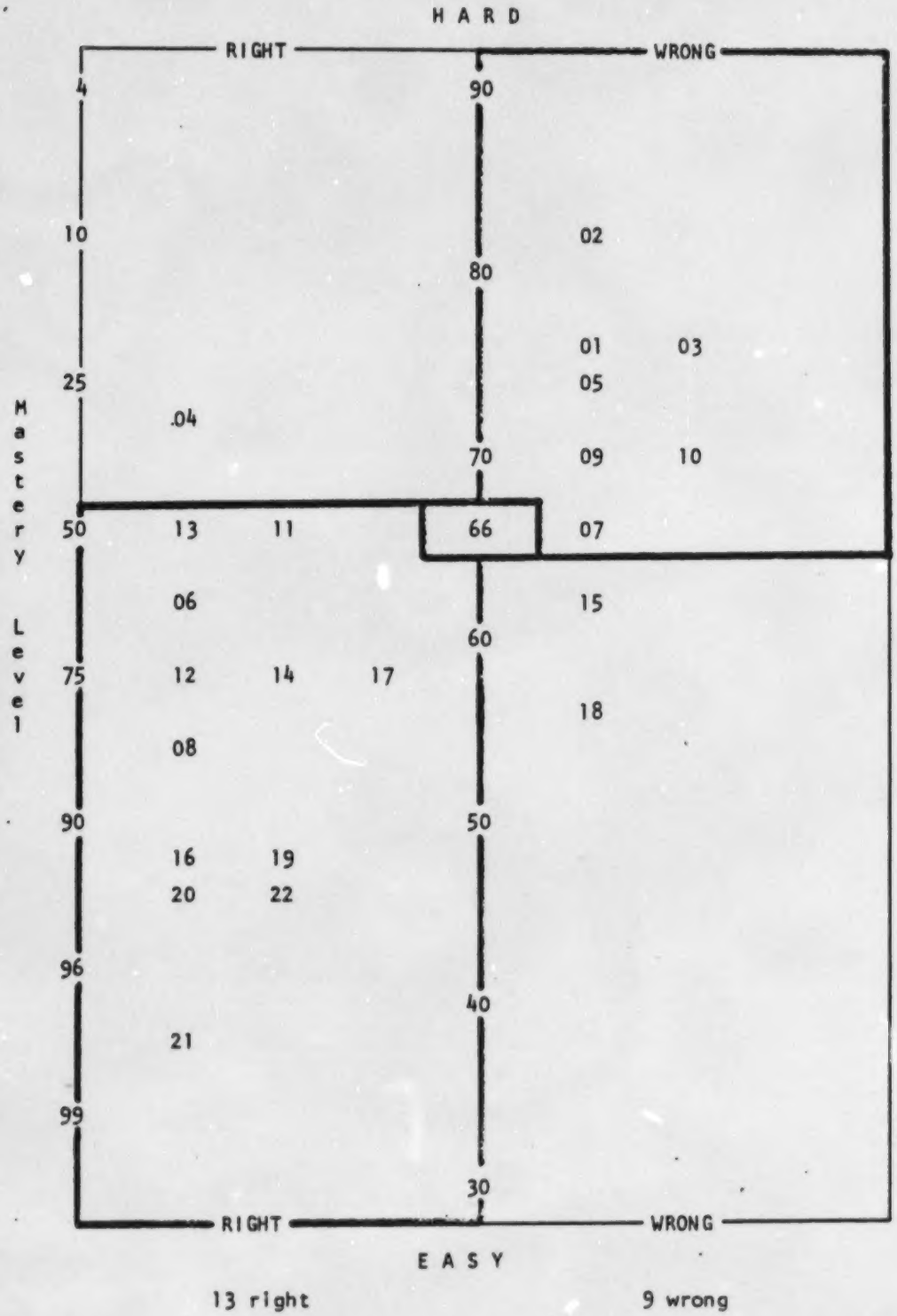
A third way in which an increasing concern with the test performances of individuals is being manifested in practice is in the development of more informative ways of reporting an individual's test performance. An example of such a report is shown in Figure 1. This particular format for reporting an individual's test performance was developed at the MESA Psychometric Laboratory at the University of Chicago (Wright, Mead and Ludlow, 1979).

Figure 1 shows the performances of one child on a 22 item science test. The child's name appears at the top left. His score of 13 on this test has been transformed to a test-free estimate (66) of his position on the science achievement variable which runs vertically up the middle of the report. The standard error of measurement (2) associated with this estimate has been used to construct a box around this child's estimated position on the science variable. The numbers in the body of the report are the item numbers. Each item is positioned vertically by its estimated difficulty. (The items become harder up the page). Items to the left of center are items which this child had right. Items to the right are items the child had wrong.

This particular child's performances are relatively orderly. He had most items below his estimated position on the variable right, and most items above his estimated position wrong. This is reflected in a high "quality" index and means that this child's estimate of 66 can be meaningfully compared to the estimates of other children whose patterns are also consistent

Name	J. SMITH
Birthday	6-9-70
Test Score	13
Measure	66
Error	2
Quality	10

Date	2-4-82
Subject	SCIENCE
Test Form	4
Test Length	22
School	HILLSIDE
Teacher	JAMES



13 right

9 wrong

with the difficulties of these items. By examining the content of the items passed and failed by this child, a content interpretation of what it means to have an achievement measure of 66 is obtained.

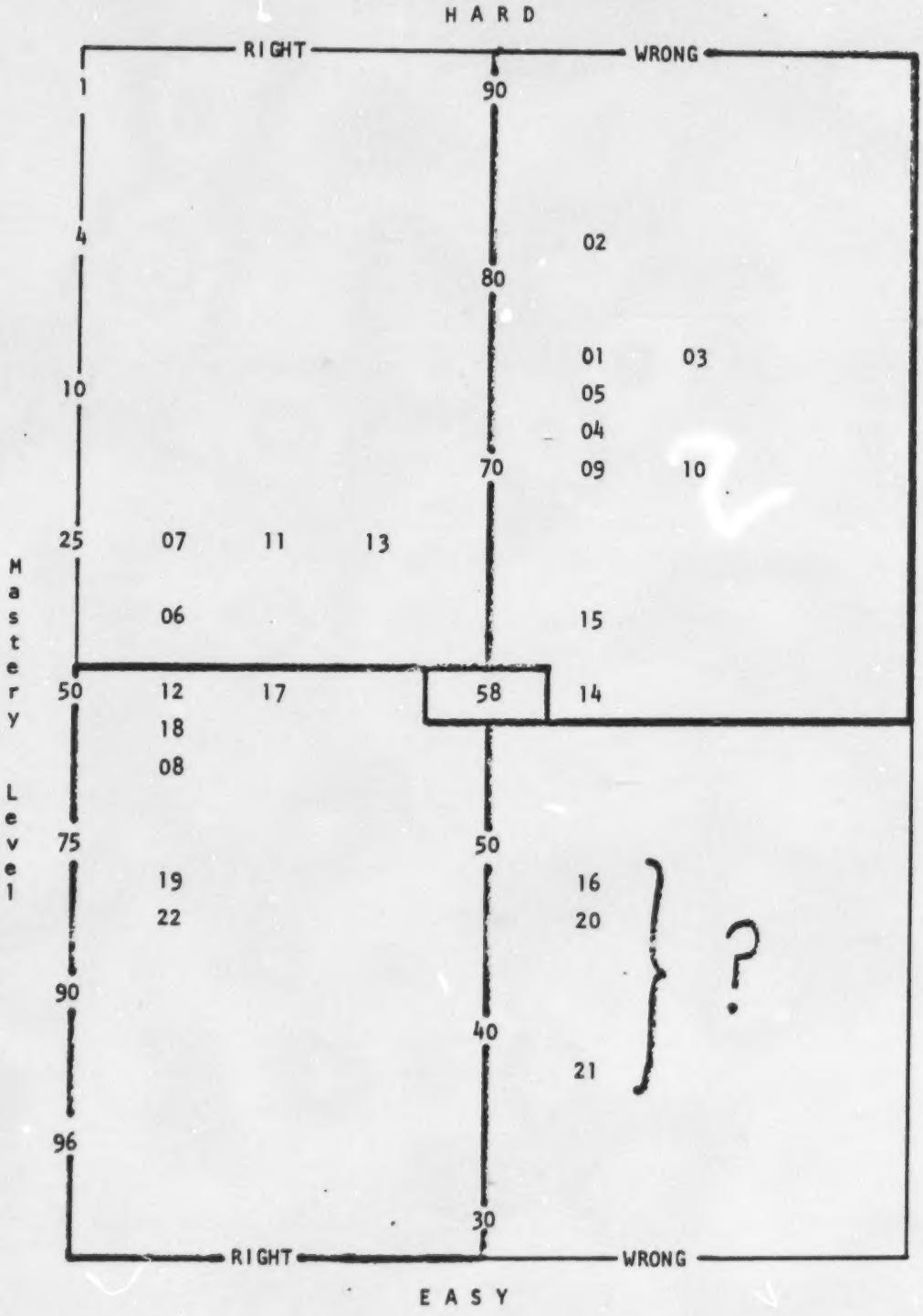
Figure 2 shows the report for another child on this same test. This child had fewer items correct, and so is estimated to be at a lower position on this science achievement variable. This child had three easy items (Items 16, 20 and 21) wrong. This is surprising for a child with an estimate of 58, and so has produced a lower "quality" index for this child. We are less certain about this child's position on this variable. An examination of easy items which are failed can provide important diagnostic information. If Items 16, 20 and 21 were all of the same type (e.g. all magnetism items) it would be important to try to understand why this child performed so poorly on these three items.

Figure 3 shows the performances of yet another child on this test. Now the surprising responses are in the upper left corner -- hard items which were answered correctly. These surprising performances raise a question about the interpretability of this child's measure of 56. Are performances on Items 2, 3 and 4 believable indications of this child's achievement or not? If not, are they the result of lucky guessing, cheating or special knowledge or coaching on some particular area of the course?

These student reports should provide students, teachers, and parents with a valuable aid to interpreting and diagnosing performances on an <u>individual</u> basis. Because they provide a

Name	A. BROWN
Birthday	19-7-70
Test Score	10
Measure	58
Error	2
Quality	04

Date	2-4-82
Subject	SCIENCE
Test Form	4
Test Length	22
School	HILLSIDE
Teacher	JAMES

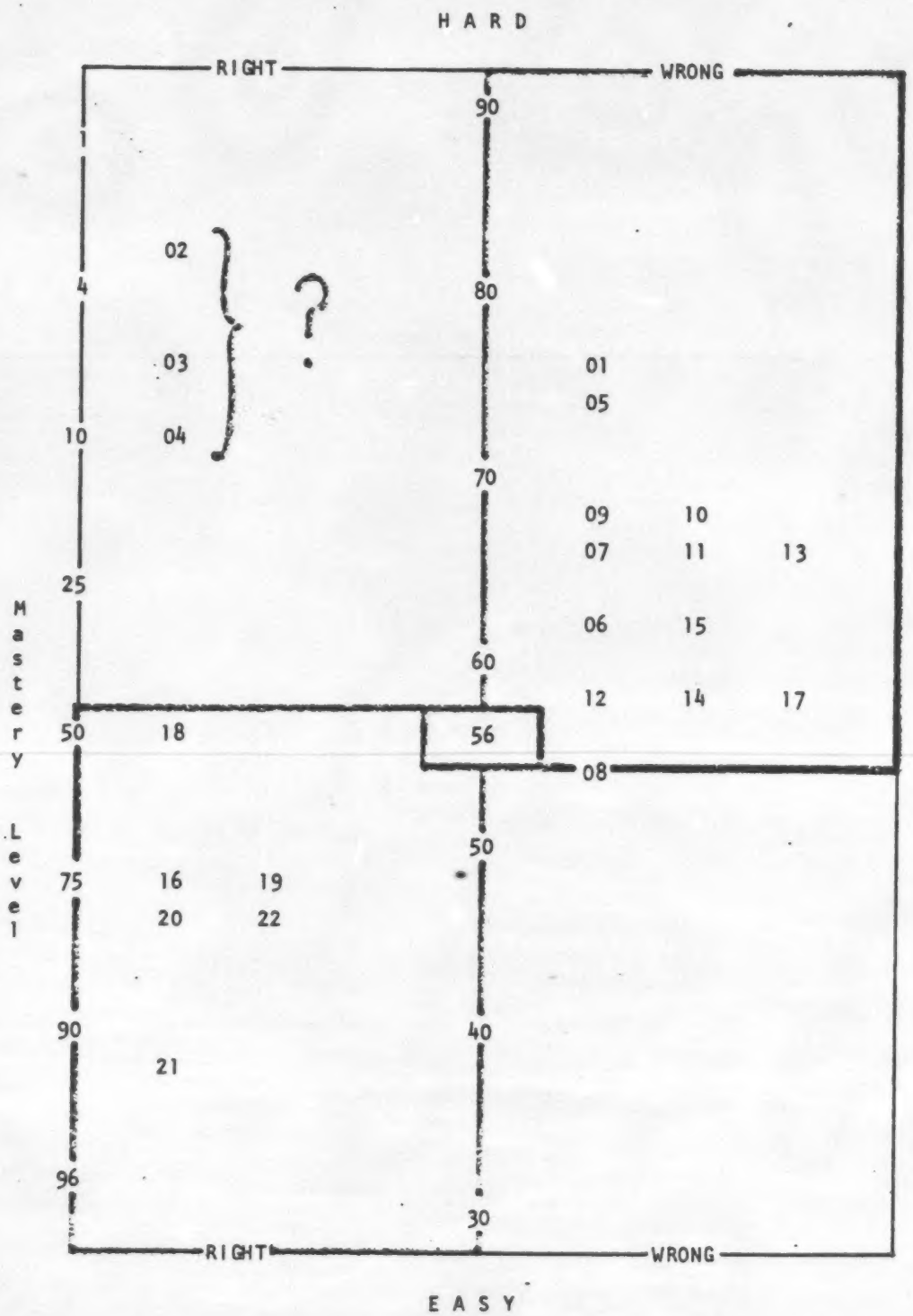


10 right

12 wrong

Name	T. JONES
Birthday	28-10-70
Test Score	9
Measure	56
Error	2
Quality	04

Date	2-4-82
Subject	SCIENCE
Test Form	4
Test Length	22
School	HILLSIDE
Teacher	JAMES



9 right

13 wrong

"test-free" estimate of a child's achievement, positions on the vertical line in the center of the report can be compared from test to test, and over time for individual students. By monitoring a student's movement up this line, his development on that particular variable can be followed.

Fairness and Bias

These student reports provide a check on the fairness of interpreting a person's score on a test as an indication of his or her achievement. An individual's surprising failures may be the result of items upon which they are disadvantaged. There are many ways in which a person may be disadvantaged on particular items. The items may require knowledge of material taught on a day when the child was absent. Or, they may cover material which the child has not had an opportunity to learn for some other reason. A child from a home in which English is not the first language, for example, will be disadvantaged on particular items which he doesn't understand or is unsure about. Similarly, a child may be disadvantaged on particular items because they require knowledge of terms which are not used in his home. (The word "sonata", for example, is a word which children with good vocabularies have often never heard.) And a child may be disadvantaged on some items because they require knowledge which they have been discouraged, because of their sex, from acquiring. A girl who scores well on a science test, for example, may fail an item about the gearbox of a car.

This approach to assessing test fairness is very different from the usual approach, which is to try to identify items which are unfair to groups of children. The traditional approach is to divide children into groups: black/white, boys/girls, and so on, and then to try to find items upon which some group tends to score poorly.

Seeing persons as members of a group and making adjustments based on group membership is inevitably unfair to those persons who are not typical of the group to which we assign them. In a sense, grouping is the start of unfairness because it denies the individuality of the persons being grouped. Most forms of discrimination and prejudice, like racism and sexism, begin with an attempt to group persons and to pretend that members of the group are identical.

When we group children and make adjustments to test scores on the basis of group membership, we inevitably give some children in the group an unfair advantage, and some children outside the group an unfair disadvantage. Test fairness is best assessed on an individual basis. Of course if sufficient individuals, for whatever reason, perform surprisingly poorly or surprisingly well on an item, then that item is best removed from our efforts to measure.

INTEGRATING LEARNING AND ASSESSMENT

A third emerging emphasis which I believe has great potential for improving the relevance of our efforts in educational testing is the trend towards making the learning process the focus of our testing efforts. Traditionally, a major role of school testing has been to facilitate administrative decision making. Terman believed that the great benefit of testing lay in "the improvement of school administration". This meant the allocation of children to streams, labelling, selection and certification. Under this view of testing, school tests have been seen primarily as conTESTS. They have been administered at the ends of courses to grade and to select. To facilitate grading and selecting, test scores have been interpreted not in terms of what children have achieved, but in terms of group norms. And, because these tests are not intended for them anyway, students have typically not seen the details of their test results or had an opportunity to examine the errors that they have made.

In contrast, I believe that in the future we can expect to see the breakdown of the distinction between assessment and learning. Testing and assessment, rather than being divorced from the learning process, will increasingly be seen as an integral and essential part of school learning. An implication of this changed emphasis will be that an increasing proportion of our testing efforts will be directed not towards the construction and analysis of summative tests, but towards the development of assessment procedures which can be used in classrooms to assess readiness and to identify gaps in a child's knowledge and understanding prior to a unit of work, to provide

feedback to the learner and to the teacher during a course of instruction so that corrective action can be taken when needed, and to assess improvement in a child's knowledge and skills as a result of a course of instruction.

When tests are used as an integral part of the learning process, there is very little place for percentiles, Z-scores, stanines and the many other notions and procedures of the norm-referenced testing tradition. Knowing how a child is achieving in relation to other children may still be of administrative use and of interest to parents, and may be important motivation for some more competitive children, but, looking ahead, I think we can expect to see tests as contests being increasingly replaced by tests as instruments for promoting school learning.

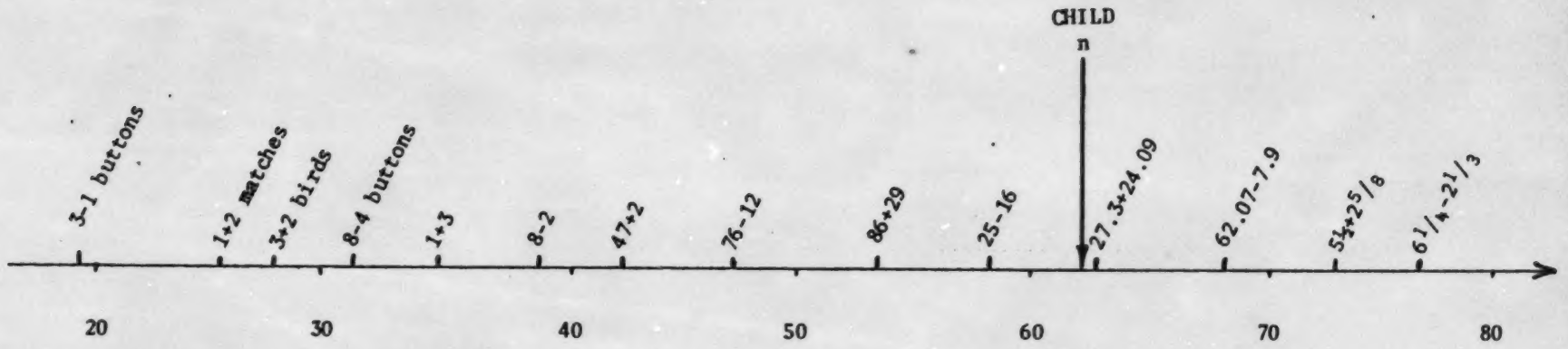
An important component in this trend towards the integration of learning and assessment is the use of tests to provide a content interpretation of a child's test performance. (I use the term "content referencing" rather than " criterion referencing", deliberately. There has been so much confusion over the definition of a "criterion-referenced" test, that the term is probably best dropped from our testing vocabulary altogether. Various definitions have equated this activity with domain referenced testing, have incorporated the notion of some sort of cut-off score, and have defined it as the interpretation of test scores with respect to performance on some external measure like job performance. In fact, the confusion over this activity has been so great, that one writer has even proposed a taxonomy for classifying definitions of criterion-referenced tests!).

Central to the notion of content referencing is the idea of a line along which persons and items have unique positions. When a person's test score provides an interpretable estimate of his position on this continuum, this estimate can be interpreted in terms of the types of items which define the continuum. In particular, the child's estimated position can be interpreted in terms of the types of tasks he has mastered, and the types of learning tasks for which he is now ready.

Figure 4 shows the estimated position of some child on a continuum of increasing mathematics achievement. His score on an arithmetic test has been interpreted as a position on this line, to the right of the easier items, and to the left of the harder items. His position estimate suggests that he has mastered the addition and subtraction of 2-digit integers. (By "mastery" we mean that he is more likely to succeed on an item requiring 2-digit addition or subtraction than to fail it.) However, he appears more likely to fail than to succeed on items requiring addition or subtraction of decimals or fractions. The implications for this child's instruction in arithmetic might be particular emphasis on the addition and subtraction of decimals, with revision of 2-digit addition and subtraction.

When the objective of testing is to provide feedback to the learner, teachers and parents, rather than to school administrators, the reporting of a test <u>score</u> or a <u>percentile</u> is not particularly informative. The interpretation of test performance in terms of the <u>content</u> of the items which define an achievement variable has much greater value as feedback about the learning process.

FIGURE 4



Based on KEYMATH Diagnostic Arithmetic Test, American Guidance Service, 1976.

It seems likely that, from the point of view of assisting in the learning process, the sooner a child receives feedback about his test performance, the better. Several procedures for providing immediate feedback to children have been proposed and are under investigation. One of these is "answer-until-correct" scoring. In answer-until-correct scoring, students record their answers to test questions by marking a special answer sheet with a pen. On a multiple-choice test, for example, a child who chooses alternative D as the correct answer to an item marks the box labelled D alongside the item. If the correct answer is chosen a ✓ appears on the answer sheet. If an incorrect alternative is marked, an X appears, and the child has the opportunity to re-read the question and to try again. The child's score is based on the number of attempts needed to provide the correct answer.

In traditional testing there is either a long delay between test administration and the return of test papers (by which time students have often lost interest anyway), or students do not see the details of their test performance at all. Answer-until-correct scoring provides immediate feedback to the learner. In this way, instead of being like talking to a brickwall, test taking becomes more of a "conversation". It gives children an opportunity to learn from their mistakes as they make them.

Some of the most promising developments for providing immediate feedback, and for the integration of testing and learning generally, are developments in the application of microcomputers to school testing. In the days when only the privileged received an education, it was relatively easy for

a child's tutor to monitor his progress and to intervene at particular points to correct misunderstandings and to provide remedial help. Once education became universal, it became necessary to teach groups of children. The task of monitoring individual progress and of providing appropriate intervention and remediation when it was needed became enormous. Under group instruction and monitoring, individual children with misunderstandings or poorly developed skills have all too frequently gone unnoticed. Intervention and corrective help have not been provided at the points in a child's schooling when they have been most needed, and so, many children have slipped further behind with each new unit of work. This has occurred in all school subjects, but it is brought to our attention most vividly when, after seven or eight years of schooling, we discover children who can barely read.

The widespread use of computers in our testing and assessment efforts is inevitable. I believe that for the assessment of at least some aspects of school learning, this tool is capable of recapturing some of the benefits of the one-on-one tutorial for monitoring the progress of individual learners. When students take tests on microcomputers, for example, the possibility exists of providing the student with immediate feedback showing whether his response to each item is right or wrong. Items of appropriate content and difficulty can be selected automatically from a bank of items, perhaps with items being selected on the basis of the student's performances on the preceding items.

At the completion of a test, a thorough analysis of each student's performance can be provided immediately. Using the bank difficulty estimates for the items taken, and the student's estimated level of achievement, attention can be drawn to surprising failures or successes. These inconsistencies in a child's test performance might be useful for diagnosing areas in which further instruction is needed. Stored information can be used to direct the learner to sources of help - particular learning activities or perhaps readings - and, when testing is complete, information about the session can be recorded for future use. This would give students an opportunity to monitor their own performance over time. Finally, a report (perhaps like the student reports described earlier) can be printed automatically for the teacher, recording the results of the testing session. This report would not only give the teacher a measure of the child's achievement, but also alert them to any areas which needed special attention.

Finally, as I have titled my talk "Looking Ahead" in educational testing and assessment, I thought I might finish with a glimpse into the future. What might an educational testing session look like by the turn of the century? Here's one possibility:

> It's 1998. Sarah Davenport, who started school in 1982, is in her final year of medical school. In this year, Sarah is expected to "further develop her patient management skills and to improve her clinical judgement as it relates to the selection of diagnostic tests, procedures, and therapeutic and surgical interventions". As part of her training she

is required to take a number of patient management problems in each of several specialist areas: gynaecology, cardiology, paediatrics, and so on.

She sits at the computer terminal in her study and logs on. She is asked to specify a specialist area, but she chooses on this occasion to be assigned a patient at random from any field. The screen clears, and she is assigned a "patient" from a bank of several thousand medical cases. A paragraph of text begins

> "A 57 year old man is brought into the emergency room complaining of severe chest pains. His wife says the pains began shortly after dinner ..."

The paragraph goes on to give a brief description of the patient as he presents.

Sarah is asked for a course of action. Using the keyboard she types in a request for a more detailed history. She follows up with several particular questions, and the answers are flashed back on to the screen. Again she is asked for a course of action. She begins monitoring vital signs. Details of heartrate, ECG, and so on appear on the screen.

She is then asked whether she wants to conduct any laboratory tests. She types in "yes" and gives the details of the tests she wants run. In time, the results of the tests come back.

A three-day health care episode is compressed into the next 30 minutes. Sarah is called on to make a variety of decisions relating to diagnosis and treatment of the simulated patient. Changes in the patient's conditions are brought to her attention automatically.

At the end of the session Sarah is complimented on several of her decisions. She is advised of actions she took which, based on expert opinion, were probably unnecessary. Any errors of judgement or omissions are pointed out, and alternative courses of action are suggested. If there appear to be gaps in her knowledge or understanding, these are pointed out, and she is referred to sources of help - medical texts, recorded medical cases or perhaps a member of faculty.

The details of the session are then stored for future reference. Based on the difficulty of the case, and on the difficulties of the decisions she faced, Sarah is given a score or perhaps a profile of scores. These enable her to compare her performance with the level of patient management skill and clinical judgement expected of a final year student, and required for certification. Details of the session are also available for access by Sarah's instructors and supervisors. These details, along with her performances on other patient management problems, will be taken into consideration in making a final decision about certification.

But for Sarah, the session is much more than a test. She sees it as part of her medical education. These sessions are an opportunity to tackle a wide variety of medical problems which she may not otherwise encounter. And she uses them to monitor her own progress in patient management.

REFERENCES

Bloom, B.S. All Our Children Learning. New York: McGraw-Hill Book Company, 1981.

Bloom, B.S. and Sosniak, L.A. Talent development vs. schooling. Educational Leadership, 1981, 39, 86-94.

Dockrell, W.B. National surveys of achievement. Issues in Educational Assessment, Scottish Education Department, Edinburgh, 1979.

Galton, F. Hereditary Genius. London: Macmillan, 1892.

Priest, T.A. (Chairman) Review of Educational Standards in Lower Secondary Schools in Western Australia. Perth: Education Department of Western Australia, 1981.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Denmarks Paedogogiske Institut, 1960 (Chicago: University of Chicago Press, 1980).

Terman, L.M. Intelligence Tests and School Reorganization. London: Harrap and Co., 1922.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. and Masters, G.N. Rating Scale Analysis. Chicago: MESA Press, 1982.

Wright, B.D., Mead, R.J. and Ludlow, L.H. KIDMAP. Research Memorandum Number 29, MESA Psychometric Laboratory, Department of Education, University of Chicago, 1980.

Thurstone, L.L. Measurement of social values. Journal of Abnormal and Social Psychology, 1931, 26, 249-269.